Research Article

Statistics in Medicine

Received 27 June 2014, Accepted 17 December 2014 Published online 1 February 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6413

Correcting for bias in the selection and validation of informative diagnostic tests

David S. Robertson,[a*†] A. Toby Prevost[b] and Jack Bowden[a]

When developing a new diagnostic test for a disease, there are often multiple candidate classifiers to choose from, and it is unclear if any will offer an improvement in performance compared with current technology. A two-stage design can be used to select a promising classifier (if one exists) in stage one for definitive validation in stage two. However, estimating the true properties of the chosen classifier is complicated by the first stage selection rules. In particular, the usual maximum likelihood estimator (MLE) that combines data from both stages will be biased high. Consequently, confidence intervals and *p*-values flowing from the MLE will also be incorrect. Building on the results of Pepe *et al.* (SIM 28:762–779), we derive the most efficient conditionally unbiased estimator and exact confidence intervals for a classifier's sensitivity in a two-stage design with arbitrary selection rules; the condition being that the trial proceeds to the validation stage. We apply our estimation strategy to data from a recent family history screening tool validation study by Walter *et al.* (BJGP 63:393–400) and are able to identify and successfully adjust for bias in the tool's estimated sensitivity to detect those at higher risk of breast cancer.

Keywords: diagnostic tests; group sequential design; family history; uniformly minimum variance unbiased estimator

1. Introduction

The development and validation of an informative diagnostic test for a medical condition is of great use for clinicians. This process is well described in the literature if only a single diagnostic variable is studied. However, there are often multiple candidate classifiers that show potential as diagnostic tools, and it may also be unclear if any will offer an improvement compared to current technology. The challenge is to identify the *most* promising diagnostic test and then to correctly validate its properties.

It is in the context of biomarker research that this challenge is particularly evident, where new technological advancements have led to an abundance of biomarker discovery studies and a huge number of candidate markers, for example, in colorectal cancer [1] and prostate cancer [2]. Guidelines have also been established for the discovery and validation of potential biomarkers [3].

The development of questionnaires for diagnosis is a parallel endeavour to biomarker discovery and validation. There will be a set of possible questions, with each considered a candidate classifier. In particular, questions about the family history of a disease are simple and cheap to measure when compared with genetic or biomarker variables. They can also provide the bulk of a diagnostic or risk prediction tool's classification ability, despite the discovery of many genetic markers [4].

To make efficient use of resources, a sequential procedure is a natural choice for the selection and validation of diagnostic tests. This is particularly the case for biomarkers, due to the high false discovery rate – despite showing initial promise, the majority of markers will not subsequently perform well enough compared with an existing test to be considered for further development. Also, many biomarker studies rely on stored biological samples, and there is a need to preserve specimen resources [5]. Hence, group sequential designs have been proposed that allow for early stopping because of poor marker performance

[a]*MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*
[b]*King's College, London, UK*
[*]*Correspondence to: David S. Robertson, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK.*
[†]*E-mail: david.robertson@mrc-bsu.cam.ac.uk*

[5, 6]. In these settings, the simplest (two-stage) group sequential design has been proposed; whereby the discovery and validation phases are separated by a single interim analysis.

Estimating the performance of the chosen classifier is complicated by the first stage selection rules. A candidate classifier will have to perform well in the first stage in order to proceed to the validation stage, which will lead to overly optimistic estimates. In particular, the usual maximum likelihood estimator (MLE) that combines data from both stages will be biased high. Hence, hypothesis-testing procedures using the MLE will have incorrect p-values, with an inflation of the type I error rate. Furthermore, confidence intervals will have coverage probabilities that can be well below the nominal level.

There are obvious parallels in this endeavour with multi-arm adaptive clinical trials of pharmaceutical treatments, where a promising single treatment or treatment dose is selected in a preliminary phase for a subsequent confirmatory analysis against standard therapy. Specific examples include seamless designs [7, 8] and drop-the-losers trials [9]. In this domain, the issues of bias and type I error inflation are well understood. Many methods exist to adjust for bias [10–13] and to ensure correct hypothesis testing [9, 14] because of demands of regulatory authorities when making licensing decisions based on trial evidence.

Bias and type I error are also important in the diagnostic test setting. Like pharmaceutical drugs, they are marketed and sold to the healthcare industry on the basis of their (claimed) clinical utility. They can have a pivotal role in guiding the treatment plan of patients [15]. Hence, diagnostic tests are subject to rigorous approval pathways by regulatory authorities.

In the spirit of Cohen and Sackrowitz [13], an efficient unbiased estimator can be obtained by taking the unbiased stage two data and conditioning on a complete, sufficient statistic – a technique known as Rao–Blackwellisation. By the Lehmann–Scheffé theorem, this will give the uniformly minimum variance conditionally unbiased estimator (UMVCUE). In a similar vein, uniformly most powerful conditionally unbiased (UMPCU) hypothesis tests have also been developed [14, 16]. The 'condition', in each case, is that the single treatment has been selected from many at stage 1 and carried forward to the validation stage.

The rationale for this continued conditional perspective is that estimation is only important if a promising classifier is actually identified. Indeed, when a study appropriately terminates early, the candidate classifiers are then deemed inadequate and further estimation of their performance is not needed. This viewpoint is demonstrated in a number of recent examples [5, 6, 17].

An alternative argument for the use of conditional estimators and confidence intervals is that we are essentially combining a discovery and validation study into a single, two-stage design. In this setting, the conditional estimators offer properties that are analogous to what would be observed if an independent validation study was completed, but are more efficient because they utilise the data from the discovery phase.

In this research article, we focus on finding the UMVCUE for the chosen classifier's sensitivity (or true positive rate) when the candidate classifiers are dichotomous. For example, this could correspond to the absence/presence of a biomarker or a 'yes'/'no' question in a questionnaire. Once the UMVCUE is found, we then construct confidence intervals for the estimated sensitivity.

Pepe *et al.* [5] considered a two-stage study for a *single* dichotomous diagnostic biomarker, with early stopping for futility. They derived the UMVCUE and described bootstrapping schemes to estimate confidence intervals for the sensitivity. Prior to this, Tappin [18] provided methodology to find the UMVCUE when selecting from multiple dichotomous classifiers (provided that ties were broken according to a pre-specified ordering) but without the option of stopping for futility or the construction of confidence intervals. This latter issue was addressed by Sill and Sampson [16], who showed how to construct *exact* confidence intervals when there are multiple candidate classifiers to choose from in the first stage.

We extend the above approaches for finding the UMVCUE and exact confidence intervals by allowing the following: (i) generalised rules for ranking the candidate classifiers; (ii) arbitrary (fixed) futility thresholds for each classifier; and (iii) unequal stage one sample sizes.

In Section 3, we describe the model framework and show how to derive the UMVCUE and construct exact confidence intervals. We then carry out a simulation study in Section 4 to investigate their properties. In Section 5, we apply our inferential technique to a recent family history screening tool validation study by Walter *et al.* [19] and conclude with a discussion in Section 6. However, we first describe the data that served as motivation for this work.

2. Motivation: The family history questionnaire study

Walter *et al.* [19] implemented a two-stage diagnostic validation study in 10 general practices across eastern England. The aim was to develop a brief self-completed family history questionnaire (FHQ) that accurately identified people at higher risk of diabetes, ischaemic heart disease (IHD), breast cancer and colorectal cancer. This self-completed FHQ would be a cheaper and simpler alternative to the current gold standard in-depth interview.

There were 1147 participants recruited into the study, with 618 in stage 1 and 529 in stage 2. This sample size was chosen to give at least 90% power to detect whether those answering 'yes' to a question would have a different risk from those answering 'no'. Overall, 32% were at an increased risk of one or more of the conditions, as assessed by the three-generational gold standard pedigree collected by trained research nurses.

In stage 1 of the analysis, the FHQ consisted of 12 questions (14 including sub-questions). Questions that were sufficiently predictive of increased risk for each condition were identified by the following procedure:

(1) Test for significance of questions using (a two-sided) Fisher's exact test with $p < 0.05$.
(2) Retain the significant question with the greatest balanced accuracy (defined as the arithmetic average of the sensitivity and specificity).
(3) Exclude each significant question if, in combination with the most accurate question, there was no significant improvement in prediction as assessed by a likelihood ratio test with $p < 0.10$.
(4) If necessary, assess further combinations of the remaining significant question using multiple logistic regression.

Questions 4a, 4b, 9a and 9b were not considered in the above analysis by Walter *et al.* because of a small number of positive responses.

Six questions (questions 2, 3, 6, 8, 10 and 11) were taken into the brief FHQ, which was tested on the additional 529 subjects in stage 2. No significant differences in sex, age or prevalence of increased risk for the conditions were found between the participants in stages 1 and 2.

Finally, to validate the retained questions, a χ^2-test was used to compare the sensitivity and specificity between the two stages for each condition. Because there were non-significant differences ($p > 0.05$) for all conditions, the data from both stages were then pooled to give an overall assessment of the brief FHQ. In particular, combined results were given for the sensitivity and specificity of the selected questions.

A schematic of the stage 1 selection process for breast cancer is given in Figure 1. Question 8 was the significant question with the highest balanced accuracy and was selected for further validation in stage 2. Question 6 was also selected on the basis of a likelihood ratio test.

Through its use of a two-stage design and a complex interim selection rule, the development of the brief FHQ has clear parallels to a biomarker discovery and validation study. Therefore, it inherits many of the same issues of bias and type I error inflation. In the next section, we describe how to derive efficient

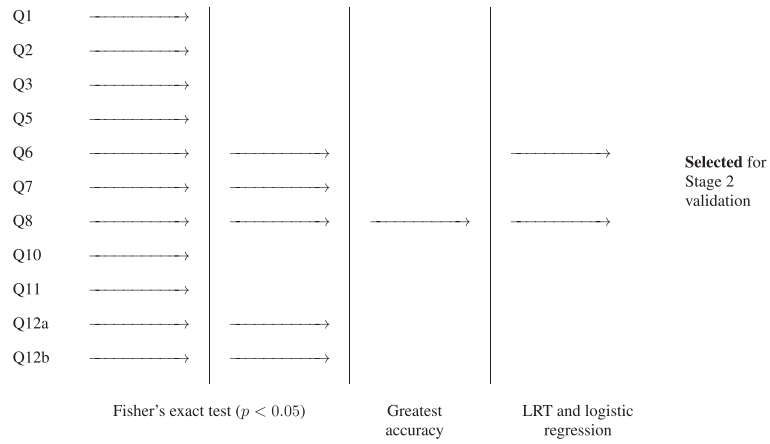

Figure 1. Schematic of the stage 1 selection process for identifying sufficiently predictive questions for breast cancer.

unbiased point estimates and confidence intervals under general selection rules, for which the FHQ study is a special case.

3. General framework for the uniformly minimum variance conditionally unbiased estimator

3.1. Model description

Suppose there are K candidate binary classifiers, each taking values in $\{0, 1\}$. For example, this could correspond to a set of K candidate diagnostic biomarkers or a questionnaire with K 'yes'/'no' questions. The aim is to select the classifier that performs 'best' (as defined below), subject to passing a 'fixed' threshold and then to estimate its sensitivity. To do so, we perform a two-stage validation study.

In the first stage, each classifier i is tested on a population that contains n_{1i} known case subjects. These could be disease cases or those that have been classified as a case by some gold standard test. Ideally, the classifiers could be all tested on the same population; hence, the n_{1i} would all be equal. However, commonly, the number of case subjects will vary between the classifiers. This could be because of missing data or because the classifier is not applicable to all subjects (e.g. gender-specific questions).

Let X_i denote the number of true positives for classifier i. Hence, we assume that we have K independent binomial variables $X_i \sim \text{Bin}\left(n_{1i}, s_i\right)$ for $i = 1, \ldots, K$, where s_i is the true *sensitivity* for the i^{th} classifier and where sensitivity is defined as Prob(positive test | subject diseased).

Each classifier has an associated fixed threshold that the number of true positives must pass in order to be considered further in stage 1. That is, for each $i \in \{1, \ldots, K\}$ there is a fixed cut-off c_i, and we require $X_i \geqslant c_i$ or else classifier i is dropped for 'futility'. For example, if there already exists a classifier with known sensitivity $\bar{c}$, then we might set $c_i = \bar{c}n_{1i}$. If all the classifiers fail to pass their respective fixed thresholds, then the whole study is terminated early.

Suppose that $L > 0$ classifiers pass their fixed threshold. Let $X_1^*, X_2^*, \ldots, X_L^*$ denote the number of true positives, where the relabelling preserves the original ordering of the labels (this is important for breaking ties). The L classifiers are then ranked from 'best' to 'worst' using a pre-specified function $r\left(X_i^*; \lambda_i\right)$, where the λ_i are constants associated with classifier i.

Thus, classifier i is ranked above classifier j if $r\left(X_i^*; \lambda_i\right) > r\left(X_j^*; \lambda_j\right)$. If there is a tie, $r\left(X_i^*; \lambda_i\right) = r\left(X_j^*; \lambda_j\right)$, we choose the classifier with the smallest index. This allows us to rank the classifiers in *a priori* order of importance. For instance, we might pre-rank the classifiers on the basis of evidence from previous studies, biological plausibility or simply the cost of measurement. A fully Bayesian approach is also possible, where classifiers are ranked using the posterior distribution of the s_i, given the specification of suitable priors. Note that the method used for breaking ties is important. For example, Tappin [18] showed that if ties are broken by randomisation, then, in fact, no UMVCUE exists.

We also require $r\left(X_i^*; \lambda_i\right)$ to induce the following inequalities on the X_i^*:

$$r\left(X_i^*; \lambda_i\right) \geqslant r\left(X_j^*; \lambda_j\right) \Rightarrow X_i^* \geqslant d\left(X_j^*; \lambda_i, \lambda_j\right) \qquad \text{for } i, j \in \{1, \ldots, L\},\ i \neq j$$

where $d\left(X_j^*; \lambda_i, \lambda_j\right)$ is a function that only depends on X_j^*, λ_i, λ_j and not on X_i^*. Hence, there is equality if and only if there is a tie in the rankings. Note that $r\left(X_i^*; \lambda_i\right)$ need not to be explicitly defined by the study organisers, as complex selection rules can be reverse engineered to conform to this set up, as we show for the FHQ study.

As an example of the above formulation, consider ranking the classifiers by their estimated sensitivities and, hence, $\lambda_i = n_{1i}$ and $r\left(X_i^*\right) = X_i^*/n_{1i}$. This induces the following inequality:

$$r\left(X_i^*; n_{1i}\right) \geqslant r\left(X_j^*; n_{1j}\right) \Rightarrow X_i^* \geqslant d\left(X_j^*; n_{1i}, n_{1j}\right) = n_{1i}X_j^*/n_{1j}.$$

At the end of stage 1, the classifier with the highest ranking (that has passed its fixed threshold) is then selected for further validation in stage 2. Let M be the index of this chosen classifier. In the second stage, the selected classifier from stage 1 is tested on a population containing n_2 additional cases, where n_2 is a constant that does not depend on X_M^*. Let Y denote the number of true positives in these n_2 additional observations. Note that $Y \sim \text{Bin}\left(n_2, s_M\right)$, independently of X_M^*.

After the end of the study, we estimate the sensitivity s_M of the selected classifier. The naïve estimator (MLE) for s_M using data from both stages is

$$\hat{S}_{\text{all}} := \frac{X_M^* + Y}{n_{1M} + n_2}.$$

This estimator is biased high, because it does not take into account the first stage selection procedures and so $E\left[X_M^*/n_{1M}|M\right] > s_M$.

An unbiased estimator $\hat{S}_2$ can easily be found by just using the stage 2 data, where $\hat{S}_2 := Y/n_2$. However, given the smaller sample size, then this estimator suffers from lower precision. Hence, we look for an unbiased estimator that utilises data from both stages.

3.2. *Deriving the uniformly minimum variance conditionally unbiased estimator*

In this section, we extend the arguments of Pepe *et al.* [5] and Tappin [18] to find the UMVCUE for the parameter of interest s_M.

Let $\left(i_1, i_2, \ldots, i_L\right)$ denote the vector of indices of the L classifiers X_i^* after they have been ranked, with ties being decided by choosing the smaller index. Hence, $M = i_1$ is the index of the selected classifier.

In what follows, we drop the * superscript for notational convenience. We drop the constants λ from the arguments of the functions r and d as well.

In Appendix A.1, we show that a complete and sufficient statistic for $\left(s_1, s_2, \ldots, s_L\right)$ is $\boldsymbol{Z} = \left(Z_1, Z_2, \ldots, Z_{2L}\right)$, where

$$Z_1 = X_{i_1} + Y, Z_2 = X_{i_2}, \ldots, Z_L = X_{i_L}$$
$$Z_{L+1} = i_1, Z_{L+2} = i_2, \ldots, Z_{2L} = i_L.$$

Let $\psi(i)$ denote the ranking of the i^{th} classifier, and Q the event

$$\left\{\psi\left(i_1\right) = 1, \psi\left(i_2\right) = 2, \ldots, \psi\left(i_L\right) = L; X_1 \geqslant c_1, X_2 \geqslant c_2, \ldots, X_L \geqslant c_L\right\}.$$

Then by the Lehmann–Scheffé theorem, $\hat{U} := E\left(\frac{Y}{n_2} \mid \boldsymbol{Z} = \boldsymbol{z}, Q\right)$ is the UMVCUE for s_M under Q.

Now, following the idea of Pepe *et al.* [5], note that conditional on $Z_1 = X_{i_1} + Y$, the distribution of Y is hypergeometric: $Y \mid Z_1 \sim \text{Hyper}\left(z_1, n_{1M} + n_2 - z_1, n_2\right)$, which can be re-expressed (for notational convenience) as $Y \mid Z_1 \sim \text{Hyper}\left(n_2, n_{1M}, z_1\right)$. That is,

$$f\left(Y|Z_1\right) = \frac{\binom{n_2}{y}\binom{n_{1M}}{z_1 - y}}{\binom{n_{1M} + n_2}{z_1}} \quad \text{for } y \in \left\{\max\left(0, z_1 - n_{1M}\right), \ldots, \min\left(z_1, n_2\right)\right\}.$$

The conditional density $f\left(Y \mid \boldsymbol{Z}, Q\right)$ is essentially the same, except that the support of Y is further restricted by Q. There is the ranking condition inequality $r\left(X_{i_1}\right) \geqslant r\left(X_{i_2}\right) \Rightarrow X_{i_1} \geqslant d\left(X_{i_2}\right)$ and the fixed threshold condition $X_{i_1} \geqslant c_{i_1}$.

The precise way that Y is additionally restricted under $(\boldsymbol{Z}, Q)$ is given below.

(1) $y + x_{i_1} = z_1$

- (a) If $i_1 > i_2 \Rightarrow$ no tie in the ranking is possible
  $\Rightarrow x_{i_1} > d\left(x_{i_2}\right) \Rightarrow y < z_1 - d(z_2)$
- (b) If $i_1 < i_2 \Rightarrow$ a tie is possible
  $\Rightarrow x_{i_1} \geqslant d\left(x_{i_2}\right) \Rightarrow y \leqslant z_1 - d\left(z_2\right)$

(2) $y + x_{i_1} = z_1$ and $x_{i_1} \geqslant c_{i_1} \Rightarrow y \leqslant z_1 - c_{i_1}$

The formula for the UMVCUE (assuming $L > 1$) is then as follows.

$$\hat{U} = E\left(\frac{Y}{n_2} \mid \boldsymbol{Z} = \boldsymbol{z}, Q\right) = \begin{cases} \dfrac{1}{n_2} \dfrac{\sum_{y \in A} y \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}}{\sum_{y \in A} \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}} & \text{if } i_1 > i_2 \text{ and } d(z_2) \in \mathbb{Z} \\ \dfrac{1}{n_2} \dfrac{\sum_{y \in B} y \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}}{\sum_{y \in B} \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}} & \text{otherwise} \end{cases} \tag{1}$$

where

$$A = \left\{\max\left(0, z_1 - n_{1M}\right), \ldots, \min\left(z_1 - d\left(z_2\right) - 1, z_1 - \lceil c_M \rceil, n_2\right)\right\} \quad : \text{conditions 1(a) and 2}$$
$$B = \left\{\max\left(0, z_1 - n_{1M}\right), \ldots, \min\left(z_1 - \left\lceil d\left(z_2\right)\right\rceil, z_1 - \lceil c_M \rceil, n_2\right)\right\} \quad : \text{conditions 1(b) and 2}$$

and $\lceil x \rceil$ is the ceiling function acting on x.

Note that if the summation over y goes up to n_2 (so either $\max(A) = n_2$ or $\max(B) = n_2$), then, in fact, $\hat{U} = \frac{z_1}{n_{1M} + n_2}$, which is just the usual MLE $\hat{S}_{\text{all}}$. This makes it clear when the stage 1 selection exerts no biasing effect at all.

If $L = 1$, then the dependence on Z_2 disappears, and we are left with the simpler formula below.

$$\hat{U} = E\left(\frac{Y}{n_2} \mid Z_1 = z_1, Q\right) = \frac{1}{n_2} \frac{\sum_{y \in A'} y \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}}{\sum_{y \in A'} \binom{n_2}{y}\binom{n_{1M}}{z_1 - y}}$$

where $A' = \left\{\max\left(0, z_1 - n_{1M}\right), \ldots, \min\left(z_1 - \lceil c_M \rceil, n_2\right)\right\}$.

3.3. *Constructing confidence intervals*

After calculating a point estimate for s_M at the end of the study, it is natural to seek a confidence interval as well. In this section, we describe two schemes for generating confidence intervals.

3.3.1. Nonparametric bootstrap. Firstly, we adapt the nonparametric bootstrap procedure originally used by Pepe *et al.* [5]. Given trial data $\boldsymbol{Z}$, the procedure follows the resampling schema below.

(1) Resample the first stage data for the selected classifier $M = i_1$ (with replacement). This gives a bootstrapped number of true positives $X_M^{(B)}$.
(2) If $X_M^{(B)} \geqslant c_M$ and $r\left(X_M^{(B)}\right) \geqslant r\left(X_{i_2}\right)$,
    (a) Resample the second stage data (with replacement), giving a bootstrapped number of true positives $Y^{(B)}$.
    (b) Calculate the UMVCUE $\hat{U}^{(B)}$ from equation (1), using $X_M^{(B)}$, $Y^{(B)}$ and the original observed value X_{i_2}.

These steps are then repeated for a large value of B, so that there are enough sampled values of $\hat{U}^{(B)}$ to accurately assess its sampling distribution. The $\alpha/2$ and $(1 - \alpha/2)$ empirical quantiles are then used as the $(1 - \alpha)\%$ confidence interval. Bootstrapped confidence intervals for the naïve estimators $\hat{S}_2$ and $\hat{S}_{\text{all}}$ are also immediately available following this procedure.

Statistics in Medicine

3.3.2. Sill–Sampson approach. Alternatively, we can adapt the approach used by Sill and Sampson [16], who found *exact* likelihood-based confidence intervals for s_M in the context of two-stage adaptive clinical trial. The derivation is similar to that in the work of Sill and Sampson [16], but we remove the control arm and also additionally allow for early stopping for futility and unequal first stage sample sizes. See Appendix A.2 for further details.

Defining $\boldsymbol{X}_{-1} := \left(X_{i_2}, \ldots, X_{i_L}\right)$, then the conditional distribution used to find the confidence intervals is

$$f_Q\left(Z_1|\boldsymbol{X}_{-1}\right) = \mu^{-1}\left[s_M/\left(1-s_M\right)\right]^{Z_1} \sum_{X_M \in D} \binom{n_{1M}}{X_M}\binom{n_2}{Z_1 - X_M}$$

where

$$\mu := \sum_{T=b}^{n_{1M}+n_2} \left[s_M/\left(1-s_M\right)\right]^{T} \sum_{X_M \in D} \binom{n_{1M}}{X_M}\binom{n_2}{T - X_M}$$

is the normalising constant and

$$D = \begin{cases} \left\{\max\left(d\left(X_{i_2}\right)+1, Z_1 - n_2, \lceil c_M \rceil, 0\right), \ldots, \min\left(Z_1, n_{1M}\right)\right\} & \text{if } i_1 > i_2 \text{ and } d\left(X_{i_2}\right) \in \mathbb{Z} \\ \left\{\max\left(\left\lceil d\left(X_{i_2}\right)\right\rceil, Z_1 - n_2, \lceil c_M \rceil, 0\right), \ldots, \min\left(Z_1, n_{1M}\right)\right\} & \text{otherwise} \end{cases}$$

$$b = \begin{cases} \max\left(d\left(X_{i_2}\right)+1, \lceil c_M \rceil, 0\right) & \text{if } i_1 > i_2 \text{ and } d\left(X_{i_2}\right) \in \mathbb{Z} \\ \max\left(\left\lceil d\left(X_{i_2}\right)\right\rceil, \lceil c_M \rceil, 0\right) & \text{otherwise.} \end{cases}$$

Suppose we observe $Z_1 = Z_{\text{obs}}$. To construct the $(1-\alpha)\%$ confidence interval for s_M, use the following functions:

$$p_1\left(s_M\right) := \sum_{Z_1=b}^{Z_{\text{obs}}} f_Q\left(Z_1|s_M, \boldsymbol{X}_{-1}\right)$$

and

$$p_2\left(s_M\right) := \sum_{Z_1=Z_{\text{obs}}}^{n_{1M}+n_2} f_Q(Z_1|s_M, \boldsymbol{X}_{-1}).$$

Bounds for a two-sided $(1-\alpha)\%$ confidence interval $\left[\Delta_1, \Delta_2\right]$ can then be found by solving the equations $p_2\left(\Delta_1\right) = \alpha_1$ and $p_1\left(\Delta_2\right) = \alpha_2$ respectively, where $\alpha_1 + \alpha_2 = \alpha$.

The original Sill–Sampson approach sets $\alpha_1 = \alpha_2 = \alpha/2$, but this does not (in general) give the shortest confidence interval. We also experimented with choosing α_1 and α_2 to minimise the confidence interval length, which we refer to as 'optimised' Sill–Sampson confidence intervals.

3.3.3. Clopper–Pearson approach. In order to see how the Sill–Sampson approach compares with using confidence intervals for the MLE, we use the well-known Clopper-Pearson method [20]. This uses the likelihood of the usual MLE to construct exact confidence intervals. Hence, the Sill–Sampson and Clopper–Pearson approaches are both likelihood based, but only the first takes into account the selection rules.

The Clopper–Pearson approach is as follows. Suppose we observe $Z_1 = Z_{\text{obs}}$. Then to construct the $(1-\alpha)\%$ confidence interval for s_M, use the following functions:

$$p_1\left(s_M\right) := \sum_{Z_1=Z_{\text{obs}}}^{n_{1M}+n_2} \binom{n_{1M}+n_2}{Z_1} s_M^{Z_1}\left(1-s_M\right)^{n_{1M}+n_2-Z_1}$$

and

$$p_2\left(s_M\right) := \sum_{Z_1=0}^{Z_{\text{obs}}} \binom{n_{1M}+n_2}{Z_1} s_M^{Z_1}\left(1-s_M\right)^{n_{1M}+n_2-Z_1}.$$

Bounds for a two-sided $(1-\alpha)\%$ confidence interval $\left[\Delta_1, \Delta_2\right]$ can then be found by solving the equations $p_2\left(\Delta_1\right) = \alpha/2$ and $p_1\left(\Delta_2\right) = \alpha/2$ respectively.

4. Simulation studies

We now perform a simulation study using a typical trial design. Consider a two-stage trial conducted on K potential diagnostic biomarkers, where the interest is in finding the biomarker with the highest sensitivity. In stage 1, the i^{th} biomarker is tested on a population that contains n_{1i} known case subjects, where the n_{1i} are not necessarily identical.

Suppose there already exists a biomarker with known sensitivity $\bar{c} = 0.70$. Hence, the fixed cut-off for biomarker i is set to $c_i = 0.70n_{1i}$. The biomarkers that satisfy $X_i \geqslant c_i$ are then ranked by sensitivity, giving $r\left(X_i\right) = X_i/n_{1i}$ and $d\left(X_j\right) = n_{1i}X_j/n_{1j}$. Finally, the selected biomarker (with label $M = i_1$) is taken forward to stage 2, where it is tested on an additional population with $n_2 = 50$ case subjects.

4.1. Point estimation

To start with, consider a simple simulation with $K = 3$ biomarkers with equal true sensitivities $\boldsymbol{S} = (0.70, 0.70, 0.70)$. Each biomarker is tested on the same population of 50 case subjects, giving $\boldsymbol{n_1} = (50, 50, 50)$ and $\boldsymbol{c} = 0.70\boldsymbol{n_1} = (35, 35, 35)$. Figure 2 gives the probability mass functions of 100,000 realisations of the three estimators $\hat{S}_{\text{all}}, \hat{S}_2$ and $\hat{U}$. Note the slight negative skew evident in the distribution of $\hat{U}$. The empirical biases and MSEs were (0.0308, -0.0001, -0.0001) and (0.0024, 0.0042, 0.0033) respectively.

Table I shows the bias and MSE of the estimators for a range of further parameter values for $\boldsymbol{S}$ and $\boldsymbol{n_1}$, where $n_2 = 50$ and $\boldsymbol{c} = 0.70\boldsymbol{n_1}$ as before. P(continue) gives the probability that the whole trial continues to the validation stage, while P(best) is the probability that the biomarker with the highest (or joint-highest) sensitivity is selected for validation in stage 2, conditional on the trial actually continuing to the validation stage.

The MLE $\hat{S}_{\text{all}}$ is biased high, and this bias is most pronounced for larger values of K and when the true sensitivities are similar. Note that $\hat{S}_{\text{all}}$ is still biased even when the probability of continuing to stage 2 is close to 100% (e.g. scenario 6). This indicates two sources of bias: the bias due to early stopping and

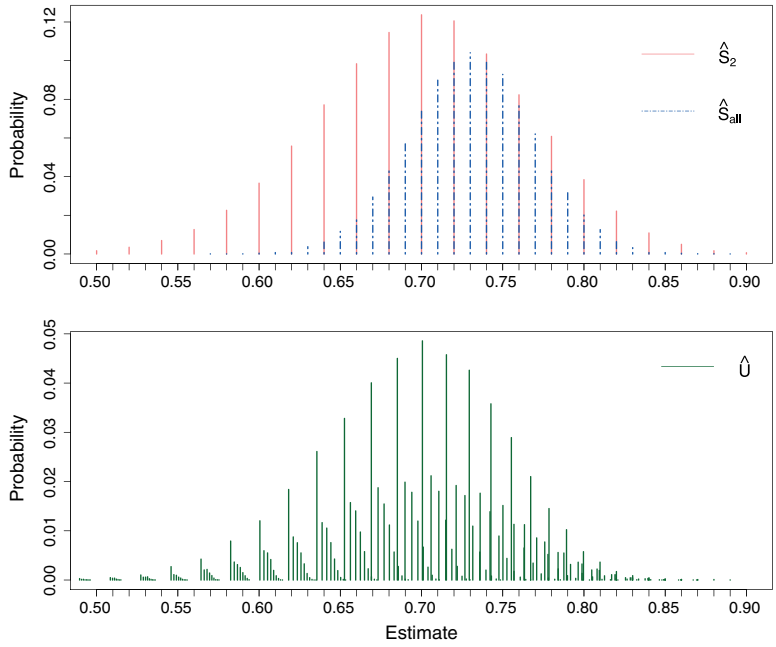

Figure 2. Probability mass functions of the estimators for $\boldsymbol{S} = (0.70, 0.70, 0.70)$, $\boldsymbol{n_1} = (50, 50, 50)$, $\boldsymbol{c} = 0.70\boldsymbol{n_1} = (35, 35, 35)$ and $n_2 = 50$. Each mass function is based on 100,000 simulations.

Table I. Simulation results with $n_2 = 50$ and $\boldsymbol{c} = 0.70\boldsymbol{n_1}$. The mean bias and MSE shown are 100 times the actual estimates. There were 100,000 simulations for each set of parameter values.

	Parameter values	P(continue)	P(best)	Bias (MSE)×100		
				$\hat{S}_{\text{all}}$	$\hat{S}_2$	$\hat{U}$
1.	$\boldsymbol{S} = (0.50, 0.70)$ $\boldsymbol{n_1} = (50, 50)$	0.570	0.997	2.289 (0.199)	0.016 (0.421)	0.000 (0.313)
2.	$\boldsymbol{S} = (0.60, 0.80)$ $\boldsymbol{n_1} = (15, 25)$	0.914	0.906	1.097 (0.222)	−0.006 (0.336)	−0.003 (0.267)
3.	$\boldsymbol{S} = (0.50, 0.70, 0.70)$ $\boldsymbol{n_1} = (25, 25, 20)$	0.810	0.987	2.909 (0.313)	−0.005 (0.420)	−0.006 (0.376)
4.	$\boldsymbol{S} = (0.50, 0.60, 0.70, 0.80)$ $\boldsymbol{n_1} = (30, 40, 40, 40)$	0.985	0.807	1.400 (0.197)	−0.015 (0.340)	−0.001 (0.244)
5.	$\boldsymbol{S} = (0.58, 0.60, 0.62, 0.64)$ $\boldsymbol{n_1} = (40, 35, 30, 30)$	0.580	0.422	4.689 (0.426)	−0.005 (0.466)	0.010 (0.418)
6.	$\boldsymbol{S} = (0.70, 0.70, 0.70, 0.70)$ $\boldsymbol{n_1} = (50, 50, 50, 50)$	0.965	1	3.465 (0.265)	0.032 (0.420)	0.023 (0.336)

the bias due to selecting the 'best' classifier from a set of candidates. The first source of bias would be expected to disappear when the probability of continuing to stage 2 is 100% but not the second.

The UMVCUE $\hat{U}$ is unbiased as expected, and it also has a lower MSE than the unbiased estimator $\hat{S}_2$ that only uses the stage 2 data. Indeed, there was a reduction of MSE ranging from 10% (for scenario 5) to 28% (for scenario 1). However, $\hat{U}$ generally has a greater MSE than $\hat{S}_{\text{all}}$, by up to 57% (for scenario 1). This is not always the case – for scenario 5, the large bias of $\hat{S}_{\text{all}}$ leads to a slightly greater MSE.

4.2. *Interval estimation*

We also consider the coverage of the confidence intervals constructed using the two procedures in Section 3.3, with $\alpha = 0.05$. Table II shows the resulting mean coverage and confidence interval width for the scenarios in Table I.

The coverage for the MLE $\hat{S}_{\text{all}}$ calculated using the nonparametric bootstrap is substantially lower than the nominal 95%, with values as low as 73% (for scenario 6). In contrast, the bootstrap coverage of the UMVCUE $\hat{U}$ is much closer to the nominal, hovering around 94% for all the scenarios. The bootstrapped confidence interval widths are greater for $\hat{U}$ than for $\hat{S}_{\text{all}}$, with an increase ranging from 16% (for scenario 2) up to 51% (for scenario 7).

Using exact (likelihood-based) approaches give better coverage for both the MLE and UMVCUE, at the cost of slightly wider confidence intervals. For the MLE, the Clopper–Pearson approach gives conservative coverage for the majority of the scenarios, except for the last two sets of parameter values where the coverage was less than the nominal 95%. In contrast, the Sill–Sampson approach gives conservative confidence intervals for all the parameter values considered. This results in an increase in confidence interval width ranging from 11% (for scenario 2) up to 31% (for scenario 7).

Using optimised Sill–Sampson confidence intervals gives a slight reduction in width and coverage, although the latter is still above 95% in all the scenarios. However, this comes at a much greater computational cost when simulating a large number of trials. Hence, we do not consider optimised Sill–Sampson confidence intervals any further in this research article.

4.3. *Hypothesis testing*

Consider now testing the hypothesis $H_0 : s_M \leqslant s_*$ versus $H_1 : s_M > s_*$, using exact 95% one-sided confidence intervals. We compare using Clopper–Pearson confidence intervals for $\hat{S}_{\text{all}}$ with the Sill–Sampson approach, where H_0 is rejected if s_* is less than the lower bound of the confidence interval. For a given set of true sensitivities $\boldsymbol{S}$, let $\boldsymbol{S_0} = \left\{s \in \boldsymbol{S} : s \leqslant s_*\right\}$. Then we define the conditional type I error rate as

Table II. Confidence interval simulation results with 10,000 simulated trials for each set of parameter values. As before, $n_2 = 50$ and $\boldsymbol{c} = 0.70\boldsymbol{n_1}$. The number of boot-strapped samples per simulation was set to $B = 10{,}000$. For comparison purposes, values for the optimised Sill–Sampson confidence intervals are shown in italics – note that only 1000 simulated trials were used here due to the computational cost.

						$\hat{\boldsymbol{S}}_{\text{all}}$			
		Sill–Sampson		$\hat{\boldsymbol{U}}$ NP bootstrap		Clopper–Pearson		NP bootstrap	
	Parameter values	Coverage	CI width	Coverage	CI width	Coverage	CI width	Coverage	CI width
1.	$\boldsymbol{S} = (0.50, 0.70)$ $\boldsymbol{n_1} = (50, 50)$	0.966 *0.956*	0.228 *0.227*	0.934	0.215	0.965	0.183	0.892	0.153
2.	$\boldsymbol{S} = (0.60, 0.80)$ $\boldsymbol{n_1} = (15, 25)$	0.969 *0.950*	0.214 *0.211*	0.945	0.196	0.966	0.193	0.938	0.169
3.	$\boldsymbol{S} = (0.70, 0.70, 0.70)$ $\boldsymbol{n_1} = (50, 50, 50)$	0.965 *0.957*	0.233 *0.232*	0.940	0.222	0.951	0.181	0.825	0.150
4.	$\boldsymbol{S} = (0.50, 0.70, 0.70)$ $\boldsymbol{n_1} = (25, 25, 20)$	0.968 *0.962*	0.249 *0.248*	0.941	0.236	0.949	0.219	0.885	0.188
5.	$\boldsymbol{S} = (0.50, 0.60, 0.70, 0.80)$ $\boldsymbol{n_1} = (30, 40, 40, 40)$	0.965 *0.964*	0.204 *0.201*	0.943	0.188	0.958	0.174	0.904	0.150
6.	$\boldsymbol{S} = (0.58, 0.60, 0.62, 0.64)$ $\boldsymbol{n_1} = (40, 35, 30, 30)$	0.961 *0.958*	0.262 *0.262*	0.943	0.256	0.913	0.212	0.731	0.180
7.	$\boldsymbol{S} = (0.70, 0.70, 0.70, 0.70)$ $\boldsymbol{n_1} = (50, 50, 50, 50)$	0.961 *0.955*	0.236 *0.235*	0.937	0.225	0.942	0.180	0.793	0.149

$\alpha = P\left(\text{reject } H_0 | s_M \in \boldsymbol{S_0}, Q\right)$. The unconditional type I error rate is defined as $P\left(\text{reject } H_0, s_M \in \boldsymbol{S_0}\right)$, where there is no conditioning on continuing to stage 2.

Similarly, the conditional power of the test is defined as $P\left(\text{reject } H_0 | s_M \in \boldsymbol{S} \backslash \boldsymbol{S_0}, Q\right)$. The unconditional power is $P\left(\text{reject } H_0, s_M \in \boldsymbol{S} \backslash \boldsymbol{S_0}\right)$, with no conditioning on continuing to stage 2.

Figure 3 shows the conditional and unconditional type I error rates and powers when the sensitivities are constrained to the set $\boldsymbol{S} = (0.50, 0.60, 0.70, 0.80)$, with stage 1 sample sizes $\boldsymbol{n_1} = (30, 40, 40, 40)$. Using the Clopper–Pearson confidence intervals for $\hat{S}_{\text{all}}$ can give highly inflated conditional type I error rates (as high as 24%), particularly for values of s_* that are just above 0.60 or 0.70.

In contrast, using the Sill–Sampson approach guarantees that the conditional type I error rate will be less than 5% for all values of s_*. This comes at the cost of lower power, both conditionally and unconditionally. Note that while using exact confidence intervals for the MLE does not control the type I error conditionally, it does control it unconditionally since $P\left(s_M \in \boldsymbol{S_0}\right)$ is low when $s_* < 0.70$.

Figure 4 shows the conditional and unconditional type I error rates and powers for the scenario $\boldsymbol{S} = (0.70, 0.70, 0.70)$ and $\boldsymbol{n_1} = (50, 50, 50)$. This time, using the confidence intervals for $\hat{S}_{\text{all}}$ gives inflated type I error rates both conditionally and unconditionally. Even unconditionally, the type I error rate can be as high as 11%. In contrast, the Sill–Sampson approach again guarantees that the type I error rate will be less than the nominal 5%. However, this is at the cost of a substantial loss of power compared with using the MLE.

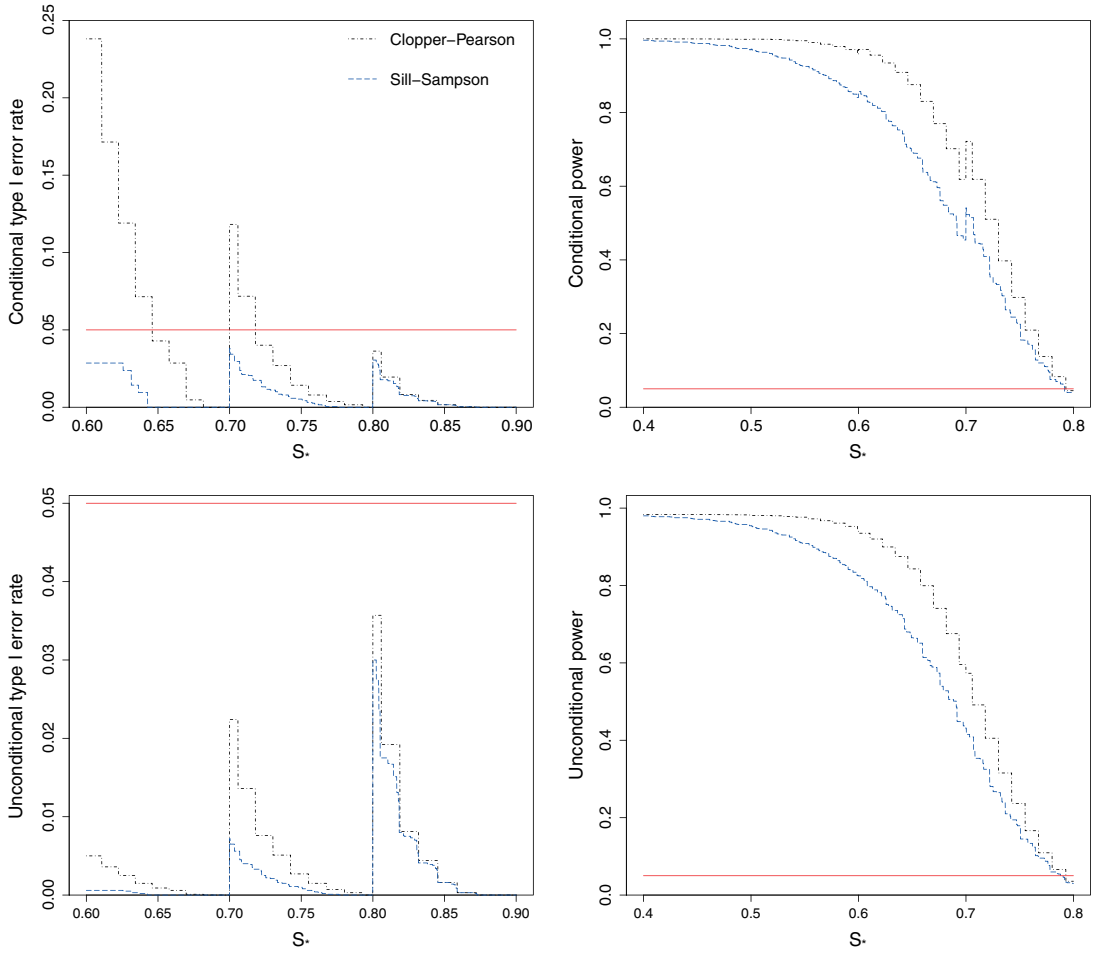

Figure 3. Conditional and unconditional type I error rates and power for testing the hypothesis $H_0 : s_M \leqslant s_*$ versus $H_1 : s_M > s_*$, using exact 95% one-sided confidence intervals. The true sensitivities are constrained to the set $\boldsymbol{S} = (0.50, 0.60, 0.70, 0.80)$, with stage 1 sample sizes $\boldsymbol{n_1} = (30, 40, 40, 40)$. Plots show the results from 10,000 simulated sets of trial data. The horizontal line shows the nominal 5% level.

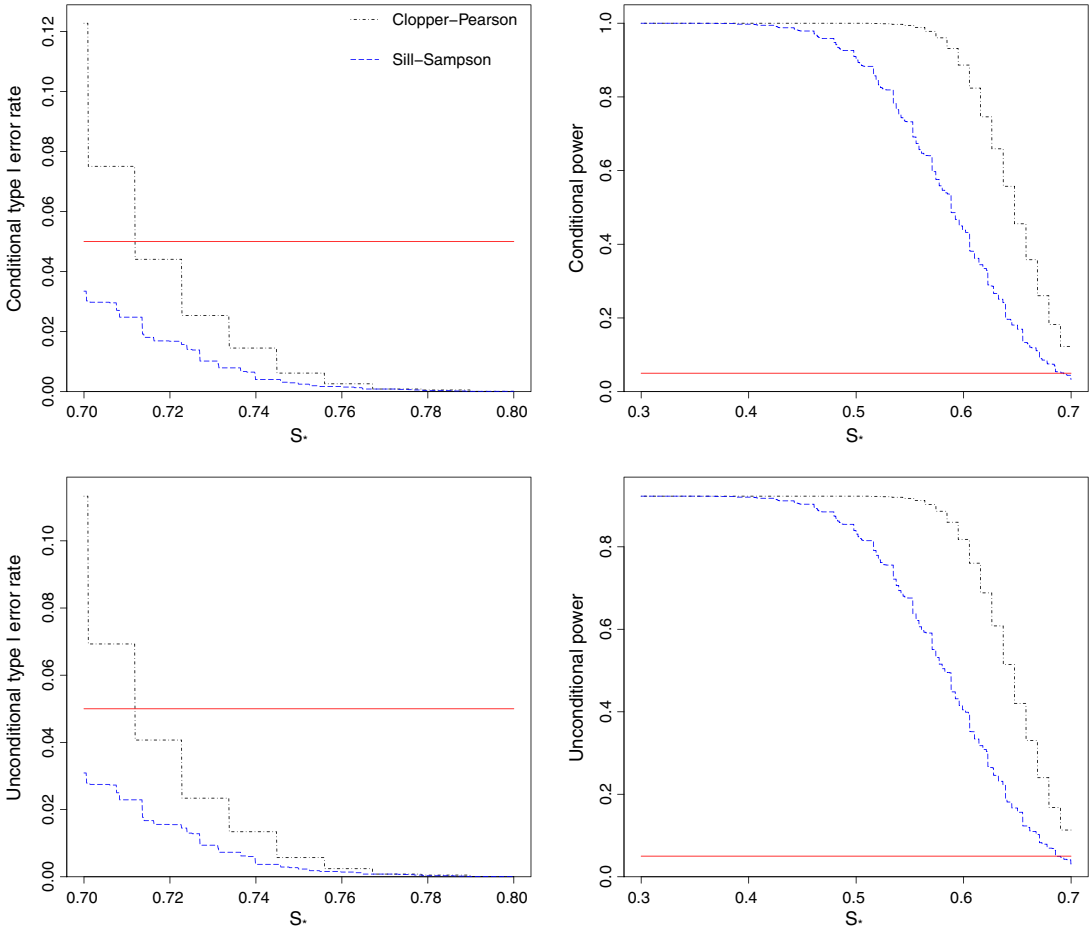

Figure 4. Conditional and unconditional type I error rates and power for testing the hypothesis $H_0 : s_M \leqslant s_*$ versus $H_1 : s_M > s_*$, using exact 95% one-sided confidence intervals. The true sensitivities are constrained to the set $\boldsymbol{S} = (0.70, 0.70, 0.70)$, with stage 1 sample sizes $\boldsymbol{n_1} = (50, 50, 50)$. Plots show the results from 10,000 simulated sets of trial data. The horizontal line shows the nominal 5% level.

5. Application to the family history questionnaire study

In this section, we return to the motivating example of the two-stage FHQ study by Walter *et al.* [19]. Although a χ^2-test for concordance was carried out before pooling data from the two stages, a natural question to ask is whether any bias was induced into the results by the stage 1 selection rules. Using the framework for bias adjusted inference outlined in Section 3, we calculate the UMVCUE and exact confidence intervals for the sensitivities of the selected questions.

5.1. Model description for the family history questionnaire

We use a slightly simplified version of the study design formulated in Section 3.1. Note that this model does not consider combinations of questions; hence, steps 3 and 4 in stage 1 are ignored. In the discussion, we comment on how the approach could potentially be extended to consider combinations of questions. In what follows, the focus is on estimating the sensitivity of the selected questions. The model for estimating the specificity, or other measures of diagnostic performance, will be very similar.

In the first stage, K questions are assessed on a case-control population, with the results for the i^{th} question available on n_{1i} cases and m_{1i} controls. Let X_i denote the number of true positives (TP) for the i^{th} question ($i = 1, \ldots, K$). That is, the total number of 'yes' responses from the case population. Then the X_i are assumed to follow independent binomial populations: $X_i \sim \text{Bin}\left(n_{1i}, s_i\right)$, where s_i denotes the true sensitivity of question i. In Section 5.3.2 we explore the performance of the method when this independence assumption in violated, as was the case (to a very limited extent) with the FHQ data.

Table III. Contingency table for Fisher's exact test.

		Question i	
		'Yes' = 1	'No' = 0
Increased risk	1	X_i	$n_{1i} - X_i$
No increased risk	0	FP_i	TNs_i

It is worth noting that when analysing the sensitivity of the selected questions, we are explicitly conditioning on the specificity results (i.e. the number of false positives and true negatives) in addition to what was specified in Section 3.2. We use this fact for both of the selection procedures: Fisher's exact test and ranking by balanced accuracy.

5.1.1. Fisher's exact test cut-off. Firstly, Fisher's exact test is applied to the contingency table given in Table III where FP_i = number of false positives and TN_i = number of true negatives for question i. As the focus is in estimating the sensitivity and we are conditioning on the observed specificity results from the trial, then the values of FP_i and TN_i are considered fixed for each i.

The aim is to find the threshold that X_i must pass in order for Fisher's exact test to give a p-value $p_i < 0.05$. That is, the value of c_i such that $X_i \geqslant c_i \Rightarrow p_i < 0.05$. Because the FP_i and TN_i are fixed, then we can do so by simply setting c_i as the smallest value in $\{0, 1, \ldots, n_{1i}\}$ such that $p_i < 0.05$ for all $X_i \geqslant c_i$.

Note that the conditioning on the observed number of false positives and true negatives is important. Indeed, another way of finding the Fisher's exact test threshold for the X_i would be to only consider the row and column totals as fixed, hence, allowing FP_i and TN_i to vary also. However, this would induce dependence between the X_i and the c_i, which would invalidate the derived form of the UMVCUE.

Although a two-sided Fisher's exact test was used in a study, we did not have to consider departures towards the other extreme – i.e. values of $X_i \ll b_i$ that gave $p_i < 0.05$. This was because all of the significant questions in the study actually passed the upper threshold c_i. In addition, we would not be interested in a question that had especially low values of X_i, because this would imply a low sensitivity. The balanced accuracy ranking (see the succeeding paragraphs) should rule out such questions being carried forward to stage 2.

In summary, for each $i \in \{1, \ldots, K\}$, there is an associated fixed threshold c_i. If $X_i \geqslant c_i$ then Fisher's exact test will give a p-value < 0.05; thus, X_i will be considered further in the balanced accuracy ranking.

Suppose $L > 0$ questions are identified as significant. Let X_i^* $(i = 1, \ldots, L)$ denote the number of true positives, where the relabelling preserves the order of the original labelling.

5.1.2. Balanced accuracy ranking. The significant question with the greatest balanced accuracy is now selected. If there is a tie (which did not occur in the study data), we assume that the question with the smallest index would be chosen.

Now, suppose question i has a greater balanced accuracy than question j. This implies the following inequality on X_i^* and X_j^*:

$$\begin{aligned} &\mathrm{Accuracy}_i \geqslant \mathrm{Accuracy}_j \\ &\quad \Rightarrow \left(\mathrm{Sensitivity}_i + \mathrm{Specificity}_i\right) \geqslant \left(\mathrm{Sensitivity}_j + \mathrm{Specificity}_j\right) \\ &\quad \Rightarrow \frac{X_i^*}{n_{1i}} + \mathrm{Sp}_i \geqslant \frac{X_j^*}{n_{1j}} + \mathrm{Sp}_j \\ &\quad \Rightarrow X_i^* \geqslant n_{1i}(\mathrm{Sp}_j - \mathrm{Sp}_i) + \frac{n_{1i}}{n_{1j}} X_j^* \end{aligned} \tag{2}$$

where $\mathrm{Sp}_i = \mathrm{Specificity}_i := \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}$.

Let $\left(i_1, i_2, \ldots, i_L\right)$ denote the vector of indices of the X_i^* after they have been ordered by balanced accuracy, and let $M = i_1$. Then from equation (2) the following inequality holds:

$$X_M^* \geqslant n_M \left(\mathrm{Sp}_{i_2} - \mathrm{Sp}_M\right) + \frac{n_{1M}}{n_{1,i_2}} X_{i_2}^*. \tag{3}$$

In the second stage, we test the selected question M from stage 1 on n_2 additional cases and m_2 additional controls. Let Y denote the number of true positives recorded in stage 2. Note that $Y \sim \text{Bin}\left(n_2, s_M\right)$, and is independent of X_M.

5.2. *The uniformly minimum variance conditionally unbiased estimator*

To find the UMVCUE for the sensitivity s_M of the selected question (after the end of stage 2), we use equation (1), where

$$d\left(Z_2\right) = \frac{n_{1M} Z_2}{n_{1,i_2}} + n_{1M}\left(\text{Sp}_M - \text{Sp}_{i_2}\right).$$

Equation (1) holds when the number of significant questions L satisfies $L > 1$, which is what occurred in this study for all of the diseases considered.

5.3. *Results*

We now apply our results to the trial data from the FHQ study, first repeating the analysis carried out in the work of Walter *et al.* [19]. Fisher's exact test indicated that an increased risk of diabetes was associated with questions 1 and 3 ($p = 0.004$ and $p < 0.001$). For IHD, questions 1, 2, 3 and 8 were significant ($p = 0.013, p < 0.001, p = 0.018$ and $p = 0.048$). For breast cancer (females only), there was a significant association for questions 6, 7, 8, 12a and 12b ($p < 0.001, p < 0.001, p < 0.001, p < 0.001$ and $p = 0.002$). Finally, increased risk of colorectal cancer was associated with questions 10 and 11 ($p < 0.001$ for both).

Table IV shows the sensitivities, Fisher's exact test thresholds (FT) and balanced accuracies for each question. The questions that passed the Fisher threshold are shown in bold, with the ultimately selected question also boxed. If the significant question with the highest balanced accuracy is chosen, then question 3 is selected for diabetes, question 2 for IHD, question 8 for breast cancer and question 10 for colorectal cancer.

5.3.1. Uniformly minimum variance conditionally unbiased estimator for the selected questions. Using the data from stages 1 and 2, we now calculate the value of the UMVCUE $\hat{U}$ for the sensitivity of the selected question for each condition, and compare it with the various naïve estimators of the sensitivity $(\hat{S}_1, \hat{S}_2, \hat{S}_{\text{all}})$. $\hat{S}_2$ and $\hat{S}_{\text{all}}$ are defined as before, while $\hat{S}_1 := X_M / n_{1M}$ is the estimated sensitivity just using the stage 1 data.

Table V gives the values of the estimators for each disease, along with exact (likelihood-based) two-sided 95% confidence intervals. For $(\hat{S}_1, \hat{S}_2, \hat{S}_{\text{all}})$, Clopper–Pearson confidence intervals are used, while the Sill–Sampson confidence interval is shown for $\hat{U}$.

For diabetes, IHD and colorectal cancer, the UMVCUE is *identical* to the MLE $\hat{S}_{\text{all}}$ that uses data from both stages. This is a consequence of the formula for $\hat{U}$ as described earlier. In addition, the Sill–Sampson confidence intervals for diabetes and IHD are virtually identical to the Clopper–Pearson intervals for $\hat{S}_{\text{all}}$. This is an attractive feature: the approach is able to identify when selection bias is *not* an issue.

However, for breast cancer, the UMVCUE is smaller than $\hat{S}_{\text{all}}$. Looking at the individual estimates for the stages $\hat{S}_1$ and $\hat{S}_2$, there is an especially large relative drop from 0.731 to 0.636 between stages 1 and 2, which supports the idea that the stage 1 data was biased high by the selection criteria. Figure 5 gives a graphical representation of the breast cancer data.

If we follow Walter *et al.* [19] and use Pearson's χ^2-test to compare the sensitivity between the two stages, the p-values are 0.873, 1.000, 0.696 and 0.920 for diabetes, IHD, breast cancer and colorectal cancer, respectively. It is interesting to note that the p-value for breast cancer is substantially lower than those for the other diseases, although it is still far above 0.05. This suggests that the χ^2-test is too conservative as a tool for detecting bias in the stage 1 data.

Indeed, for breast cancer, suppose we assume that the stage 1 data as well as the total number of cases in stage 2 are fixed. Then the number of true positives in stage 2 would have to be less than or equal to 8 (i.e. a sensitivity less than 0.363) in order for the χ^2-test to reject the null hypothesis.

5.3.2. Correlation. Finally, we consider the effect of correlation on the sensitivity estimates for the FHQ data. Recall that the data were assumed to be drawn from independent populations. However, in the FHQ study, each participant answered multiple questions. It sometimes happens that the answer to

Table IV. Stage 1 sensitivities, Fisher's exact test thresholds (FT) and balanced accuracies for each condition. Questions passing the FT are shown in bold, with the selected question also boxed.

	Diabetes			IHD			Breast cancer			Colorectal cancer		
Q.	Sensitivity	FT	Accuracy	Sensitivity	FT	Accuracy	Sensitivity	FT	Accuracy	Sensitivity	FT	Accuracy
1	**91/114**	**87/114**	**0.570**	**64/79**	**62/79**	**0.570**	23/26	24/26	0.577	10/13	13/13	0.542
2	42/113	44/113	0.544	**74/80**	**25/80**	**0.859**	9/26	13/26	0.521	2/13	8/13	0.425
3	**112/114**	**14/114**	**0.959**	**29/80**	**28/80**	**0.566**	6/26	12/26	0.488	5/13	7/13	0.570
5	4/113	9/113	0.500	3/80	7/80	0.502	1/25	4/25	0.504	0/13	3/13	0.482
6	6/113	11/113	0.506	5/80	8/80	0.512	**7/25**	**3/25**	**0.626**	0/13	3/13	0.478
7	11/114	15/114	0.515	6/80	12/80	0.501	**13/26**	**4/26**	**0.727**	1/13	4/13	0.502
8	20/113	27/113	0.511	**19/78**	**19/78**	**0.546**	**19/26**	**9/26**	**0.783**	2/13	6/13	0.496
10	7/113	14/113	0.499	4/76	11/76	0.494	2/26	4/26	0.514	**11/13**	**3/13**	**0.901**
11	18/113	21/113	0.526	12/77	16/77	0.522	3/26	8/26	0.498	**7/13**	**4/13**	**0.714**
12a	82/113	90/113	0.516	53/79	65/79	0.484	**26/26**	**23/26**	**0.657**	10/13	13/13	0.536
12b	33/107	41/107	0.512	23/74	30/74	0.513	**14/24**	**12/24**	**0.656**	3/12	7/12	0.481

Table V. Uniformly minimum variance conditionally unbiased estimators (UMVCUE) and naïve estimators for the selected questions for each disease, with exact (likelihood-based) 95% confidence intervals.

Condition	Question	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_{all}$	$\hat{U}$
Diabetes	3	0.982 (0.938, 0.998)	0.970 (0.914, 0.994)	0.977 (0.946, 0.992)	0.977 (0.946, 0.992)
Ischaemic heart disease	2	0.925 (0.844, 0.972)	0.931 (0.845, 0.977)	0.928 (0.874, 0.963)	0.928 (0.874, 0.963)
Breast cancer	8	0.731 (0.522, 0.884)	0.636 (0.407, 0.828)	0.688 (0.537, 0.813)	0.662 (0.455, 0.806)
Colorectal cancer	10	0.846 (0.546, 0.981)	0.750 (0.428, 0.945)	0.800 (0.593, 0.932)	0.800 (0.579, 0.932)

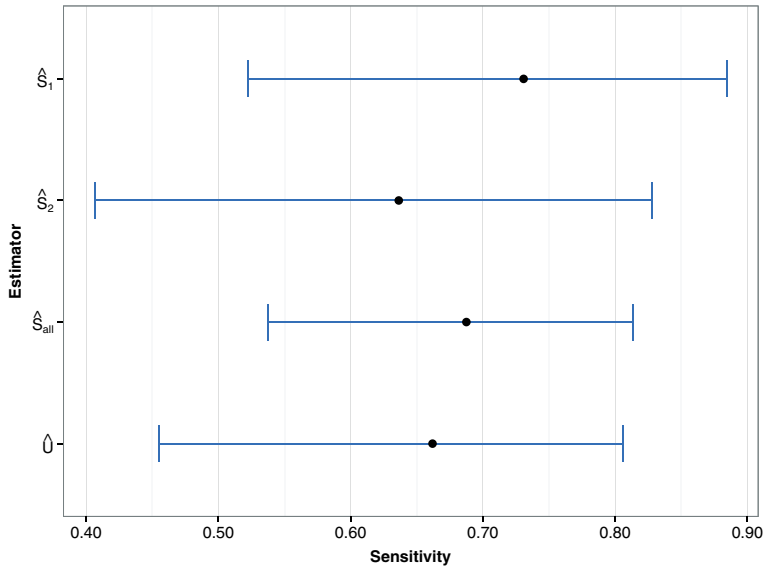

Figure 5. Plot of point estimates and exact (likelihood-based) 95% confidence intervals for the breast cancer data.

Table VI. Correlation matrix for all of the stage 1 data in the family history questionnaire study, using pairwise-complete responses.

	Q1	Q2	Q3	Q5	Q6	Q7	Q8	Q10	Q11	Q12a	Q12b
Q1	1	0.142	0.121	0.051	0.072	0.028	0.096	0.047	0.086	0.049	0.062
Q2	0.142	1	0.101	0.090	0.056	0.033	0.062	0.006	0.050	0.002	0.063
Q3	0.121	0.101	1	−0.003	0.067	0.027	0.013	0.005	0.069	0.022	0.029
Q5	0.051	0.090	−0.003	1	0.006	0.016	0.014	−0.049	−0.014	0.121	0.012
Q6	0.072	0.056	0.067	0.006	1	0.005	0.018	0.045	−0.002	0.085	0.096
Q7	0.028	0.033	0.027	0.016	0.005	1	0.282	0.004	−0.007	0.116	0.124
Q8	0.096	0.062	0.013	0.014	0.018	0.282	1	−0.023	0.056	0.240	0.112
Q10	0.047	0.006	0.005	−0.049	0.045	0.004	−0.023	1	0.193	0.084	0.112
Q11	0.086	0.050	0.069	−0.014	−0.002	−0.007	0.056	0.193	1	0.187	0.121
Q12a	0.049	0.002	0.022	0.121	0.085	0.116	0.240	0.084	0.187	1	0.395
Q12b	0.062	0.063	0.029	0.012	0.096	0.124	0.112	0.112	0.121	0.395	1

one questions should (logically at least) determine the answer to another question as well. For example, answering ‘yes’ to question 12b should mean that the answer to question 12a will also be ‘yes’. For these two reasons, we might expect there to be some correlation between the sensitivity estimates for different questions. The correlation matrix (using pairwise-complete responses) for all of the stage 1 data is displayed in Table VI.

Reassuringly, the correlations between all of the questions appears to be rather small, with a mean (absolute) pairwise correlation of just 0.07. The maximum correlation coefficient was 0.395, for the pair (Q12a, Q12b), which is explained by the aforementioned reason.

Nevertheless, we simulated FHQ-like data with the above correlation structure, using a modified version of the R package `bindata` [21]. The true sensitivities were assumed equal to the estimated stage 1 sensitivities, with 50,000 simulated data sets for each condition. For breast cancer the UMVCUE had a mean bias of -0.0082, which is less than 32% of the observed correction to the MLE for the actual FHQ data. For the other diseases there was no appreciable bias.

6. Discussion

In this research article, we present a framework for conditional estimation for a general two-stage trial design with binary classifiers. By allowing for generalised selection rules and arbitrary futility thresholds, our estimation strategy can be applied to a wide range of two-stage validation study designs. In particular, complex ranking criteria can be reverse engineered to fit within our framework.

We showed that using the usual MLE can lead to substantial conditional bias, especially when there are many candidate classifiers under consideration with similar true sensitivities. In contrast, the UMVCUE is indeed unbiased but often at the expense of a larger MSE. However, there are still large savings in efficiency when compared with just using the unbiased stage 2 data.

The usual MLE also can suffer from incorrect confidence interval coverage and inflated type I error rates for hypothesis testing, both conditionally and unconditionally. These issues can be avoided by using the Sill–Sampson approach to find exact confidence intervals, although this comes at the cost of reduced power. Although this approach is somewhat conservative, when presenting the results of a trial to a regulatory authority, any inflation in the type I error rate above the advertised level is likely to be deemed unacceptable [16].

The application of our inferential technique to the FHQ data demonstrated how the UMVCUE can identify whether selection bias is an issue. Point estimates for the selected questions using the UMVCUE and the MLE were identical for three of the conditions, with virtually identical confidence intervals as well. However, for breast cancer, the UMVCUE was able to identify and correct for the bias induced in the MLE. We also found that with the correlation structure present in the FHQ data, these results were not significantly affected by the minor violations of the independence assumption.

Our focus in this research article was in deriving unbiased estimators for the true sensitivity of the chosen classifier. However, by relaxing the unbiasedness condition slightly, it may be possible to achieve a lower MSE. One approach we tried was to use *median unbiased estimates*, as described by Jovic and Whitehead [22]. Briefly, using the distribution functions $p_1\left(s_M\right)$ and $p_2\left(s_M\right)$ defined for the Sill–Sampson approach, the (approximate) median unbiased estimator is given by $\frac{1}{2}\left(\Delta_1+\Delta_2\right)$, where $p_2\left(\Delta_1\right)=0.5$ and $p_1\left(\Delta_2\right)=0.5$. However, we found that there was no gain over the UMVCUE in terms of MSE, and the estimator was indeed biased slightly low in its mean.

We only considered a design that selects and evaluates the performance of a single classifier. However, many studies (including the FHQ study) *combine* multiple classifiers into *risk prediction models*. Much further research is needed to explore conditional estimation for combinations of classifiers, especially given the wide variety of model selection and validation procedures present in the literature. For example, recent work by Koopmeiners *et al.* [23] describes the issue of testing and validating a panel of biomarkers. Accounting for correlation will clearly be essential here too.

One way to try and deal with correlated classifiers is to decorrelate the variables of interest, as described by Zuber and Strimmer [24] in the context of biomarker discovery and gene-ranking by *t*-scores. However, is not clear whether similar transformations can be applied to binary data without altering its distribution.

A related issue would be to consider *joint* inference on the sensitivity and specificity. As mentioned in Section 5, by conditioning on the observed specificity results we treated the number of true negatives (and false positives) as fixed values. If we instead considered the number of true negatives as a binomial random variable (possibly correlated with the number of true positives), then further work would be needed to allow conditionally unbiased estimation. A complicating factor would be determining how the ranking criterion and 'fixed' thresholds change as the number of true positives and negatives are jointly varied.

Finally, another extension would be to consider inference trials with more than two stages. Bowden and Glimm [11] describe conditionally unbiased estimates for normally distributed outcomes with multiple stages of selection, and their approach could be extended to the binomial setting.

Appendix A

A.1. Proof of the completeness and sufficiency of Z

Here we prove the following theorems (originally theorems 2.1 and 3.1 in the work of Tappin [18]).

Theorem A.1
The statistic $\boldsymbol{Z} = \left(Z_1, Z_2, \ldots, Z_{2L}\right)$ is sufficient for $\left(s_1, s_2, \ldots, s_L\right)$, where

$$Z_1 = X_M + Y, Z_2 = X_{i_2}, \ldots, Z_L = X_{i_L}$$
$$Z_{L+1} = i_1, Z_{L+2} = i_2, \ldots, Z_{2L} = i_L$$

Proof
The joint distribution of $\boldsymbol{X}, Y$ is as follows:

$$\begin{aligned} f\left(\boldsymbol{X}, Y\right) &= \binom{n_{1M}}{X_M} s_M^{X_M} \left(1 - s_M\right)^{n_{1M} - X_M} \binom{n_2}{Y} s_M^{Y} \left(1 - s_M\right)^{n_2 - Y} \\ &\quad \times \prod_{j \neq M} \binom{n_{1j}}{X_j} s_j^{X_j} \left(1 - s_j\right)^{n_{1j} - X_j} \\ &= \left[s_M^{X_M + Y} \left(1 - s_M\right)^{n_{1M} + n_2 - \left(X_M + Y\right)} \prod_{j \neq M} s_j^{X_j} \left(1 - s_j\right)^{n_{1j} - X_j} \right] \\ &\quad \times \left[\binom{n_{1M}}{X_M} \binom{n_2}{Y} \prod_{j \neq M} \binom{n_{1j}}{X_j} \right]. \end{aligned}$$

Thus according to the factorisation criteria, the statistic $\boldsymbol{Z}$ is sufficient for $\left(s_1, s_2, \ldots, s_L\right)$. □

Theorem A.2
When ties are broken by selecting the population with the smallest index, the sufficient statistic $\boldsymbol{Z}$ is also complete.

Proof
Following Tappin [18] and Jung and Kim [25], we prove the result for $L = 2$ classifiers, and note that the argument easily extends to arbitrary $L > 2$.

For an arbitrary function $g(\cdot)$, defined on the range of $\boldsymbol{Z}$, we will show that $E_{\boldsymbol{s}}\left[g(\boldsymbol{z})\right] \equiv 0$ for all $\boldsymbol{s} \implies g(\boldsymbol{z}) \equiv 0$. Without loss of generality, we assume that $d\left(z_2\right) \in \mathbb{Z}$. Let

$$A_1 = \left\{\boldsymbol{Z} : z_3 = 1, \lceil c_{i_2} \rceil \leqslant z_2 \leqslant n_{1i_2}, \max\left(0, \lceil c_M \rceil, d\left(z_2\right)\right) \leqslant z_1 \leqslant n_{1M} + n_2\right\}$$
$$A_2 = \left\{\boldsymbol{Z} : z_3 = 2, \lceil c_{i_2} \rceil \leqslant z_2 \leqslant n_{1i_2}, \max\left(0, \lceil c_M \rceil, d\left(z_2\right) + 1\right) \leqslant z_1 \leqslant n_{1M} + n_2\right\}.$$

Note that $Z_1 = X_M, Z_2 = X_{i_2}, Z_3 = M = i_1$ and $Z_4 = i_2$. Because $Z_1 = X_M + Y$, then the distribution of $\left(\boldsymbol{X}, Z_1\right)$ is given by

$$f\left(\boldsymbol{X}, Z_1\right) = \binom{n_{1M}}{X_M} \binom{n_{12}}{Z_2} \binom{n_2}{Z_1 - X_M} s_M^{Z_1} \left(1 - s_M\right)^{n_{1M} + n_2 - Z_1} s_2^{Z_2} \left(1 - s_2\right)^{n_{12} - Z_2}$$

We can now find the distribution of $\boldsymbol{Z}$ by summing over the support of X_M:

$$f(\boldsymbol{Z}) = \kappa(\boldsymbol{z}) s_M^{Z_1} \left(1 - s_M\right)^{n_{1M} + n_2 - Z_1} s_2^{Z_2} \left(1 - s_2\right)^{n_{12} - Z_2}.$$

where

$$\kappa(z) = \left[\binom{n_{12}}{Z_2} \sum_{X_M \in D} \binom{n_{1M}}{X_M} \binom{n_2}{Z_1 - X_M} \right]$$

and

$$D = \begin{cases} \{\max(d(Z_2) + 1, Z_1 - n_2, \lceil c_M \rceil, 0), \ldots, \min(Z_1, n_{1M})\} & \text{if } i_1 > i_2 \\ \{\max(\lceil d(Z_2) \rceil, Z_1 - n_2, \lceil c_M \rceil, 0), \ldots, \min(Z_1, n_{1M})\} & \text{otherwise.} \end{cases}$$

Thus

$$\begin{aligned} h(\boldsymbol{s}) := E_{\boldsymbol{s}}\left[g(\boldsymbol{z})\right] = & \sum_{z \in A_1} g(\boldsymbol{z})\kappa(\boldsymbol{z}) \left(1 - s_1\right)^{n_{11} + n_2 - Z_1} s_2^{Z_2} \left(1 - s_2\right)^{n_{12} - Z_2} \\ & + \sum_{z \in A_2} g(\boldsymbol{z})\kappa(\boldsymbol{z}) s_2^{Z_1} \left(1 - s_2\right)^{n_{12} + n_2 - Z_1} s_1^{Z_2} \left(1 - s_1\right)^{n_{1i_2} - Z_2} \end{aligned} \tag{4}$$

Let $P(\boldsymbol{s}, j, k) := h(\boldsymbol{s})/s_j^k$ and $Q(\boldsymbol{s}, j, l) := h(\boldsymbol{s})/(1 - s_j)^l$ for $j \in \{1, 2\}$. Each term, say term i, in equation (4) has the factor $s_1^{k_{1i}} \left(1 - s_1\right)^{l_{1i}} s_2^{k_{2i}} \left(1 - s_2\right)^{l_{2i}}$ for some non-negative integers $k_{1i}, k_{2i}, l_{1i}, l_{2i}$. Because all terms have different factors, that is, $\left(k_{1i}, k_{2i}, l_{1i}, l_{2i}\right) \neq \left(k_{1j}, k_{2j}, l_{1j}, l_{2j}\right)$ for $i \neq j$, any subset of the terms in equation (4) has a unique minimum among $\left\{k_{1i}, k_{2i}, l_{1i}, l_{2i}\right\}$.

On the one hand, if $\left\{k_{1i}\right\}$ has a unique minimum k_1 and because $P\left(\boldsymbol{s}, 1, k_1\right) = 0$ for all $\boldsymbol{s}$, letting $s_1 \to 0$ and $s_2 > 0$ show that $g(\boldsymbol{z}) = 0$, where $g(\boldsymbol{z})$ is the coefficient of the term with the $s_1^{k_1}$ factor. Similarly, if $\left\{k_{2i}\right\}$ has a unique minimum k_2 and because $P\left(\boldsymbol{s}, 2, k_2\right) = 0$ for all $\boldsymbol{s}$, letting $s_2 \to 0$ and $s_1 > 0$ show that $g(\boldsymbol{z}) = 0$, where $g(\boldsymbol{z})$ is the coefficient of the term with the $s_2^{k_2}$ factor.

On the other hand, if $\left\{l_{1i}\right\}$ has a unique minimum l_1 and because $Q\left(\boldsymbol{s}, 1, l_1\right) = 0$ for all $\boldsymbol{s}$, letting $s_1 \to 1$ and $s_2 > 0$ show $g(\boldsymbol{z}) = 0$, where $g(\boldsymbol{z})$ is the coefficient of the term with the $\left(1 - s_1\right)^{l_1}$ factor. Similarly, if $\left\{l_{2i}\right\}$ has a unique minimum l_2 and because $Q\left(\boldsymbol{s}, 2, l_2\right) = 0$ for all $\boldsymbol{s}$, letting $s_2 \to 1$ and $s_1 > 0$ show $g(\boldsymbol{z}) = 0$, where $g(\boldsymbol{z})$ is the coefficient of the term with the $\left(1 - s_2\right)^{l_2}$ factor.

Whichever coefficient is 0, we remove that term from $h(\boldsymbol{s})$ before the next step. We continue this procedure until all terms in equation (4) are removed, concluding that $g(\boldsymbol{z}) \equiv 0$ for all $\boldsymbol{z}$ in the support of $\boldsymbol{Z}$. □

A.2. Derivation of the Sill–Sampson approach

Defining $\boldsymbol{X} = \left(X_1, \ldots, X_L\right)$, consider the joint distribution of $\boldsymbol{X}, Y|Q$:

$$f_Q(\boldsymbol{X}, Y) = K(\boldsymbol{s})^{-1} I_Q(\boldsymbol{X}) \Pi \binom{n_{1M}}{X_M} s_M^{X_M} \left(1 - s_M\right)^{n_{1M} - X_M} \binom{n_2}{Y} s_M^Y \left(1 - s_M\right)^{n_2 - Y}$$

where $K(\boldsymbol{s})$, with $\boldsymbol{s} = \left(s_1, \ldots, s_L\right)$, is the probability of observing the event Q, $I_Q(\boldsymbol{X})$ is the indicator function for Q, and

$$\Pi = \prod_{j \neq M} \binom{n_{1j}}{X_j} s_j^{X_j} \left(1 - s_j\right)^{n_{1j} - X_j}$$

Because $Z_1 = X_M + Y$, the distribution of $\boldsymbol{X}, Z_1|Q$ is given by

$$f_Q\left(\boldsymbol{X}, Z_1\right) = K(\boldsymbol{s})^{-1} I_Q(\boldsymbol{X}) \Pi \binom{n_{1M}}{X_M} \binom{n_2}{Z_1 - X_M} s_M^{Z_1} \left(1 - s_M\right)^{n_{1M} + n_2 - Z_1}$$

We can now find the distribution of the complete sufficient statistic $\boldsymbol{Z}$ conditional on Q by summing over the support of X_M:

$$f_Q(\boldsymbol{Z}) = f_Q\left(Z_1, \boldsymbol{X}_{-1}\right) = \Pi K(\boldsymbol{s})^{-1} I_{Q'}\left(\boldsymbol{X}_{-1}\right) s_M^{Z_1} \left(1 - s_M\right)^{n_{1M}+n_2-Z_1} \sum_{X_M \in D} \binom{n_{1M}}{X_M} \binom{n_2}{Z_1 - X_M}.$$

where $I_{Q'}\left(\boldsymbol{X}_{-1}\right)$ is the indicator function for $\boldsymbol{X}_{-1} := \left(X_{i_2}, \ldots, X_{i_L}\right)$ on $Q' = \left(\psi(i_2) = 2, \ldots, \psi\left(i_L\right) = L; X_{i_2} \geqslant c_{i_2}, \ldots, X_{i_L} \geqslant c_{i_L}\right)$ and

$$D = \begin{cases} \left\{\max\left(d\left(Z_2\right) + 1, Z_1 - n_2, \lceil c_M \rceil, 0\right), \ldots, \min(Z_1, n_{1M})\right\} & \text{if } i_1 > i_2 \text{ and } d(Z_2) \in \mathbb{Z} \\ \left\{\max\left(\left\lceil d\left(Z_2\right)\right\rceil, Z_1 - n_2, \lceil c_M \rceil, 0\right), \ldots, \min(Z_1, n_{1M})\right\} & \text{otherwise.} \end{cases}$$

Then the distribution of $\boldsymbol{X}_{-1}$ is

$$f_Q\left(\boldsymbol{X}_{-1}\right) = \sum_{Z_1=b}^{n_{1M}+n_2} f_Q\left(Z_1, \boldsymbol{X}_{-1}\right)$$

where

$$b = \begin{cases} \max\left(d\left(Z_2\right) + 1, \lceil c_M \rceil, 0\right) & \text{if } i_1 > i_2 \text{ and } d\left(z_2\right) \in \mathbb{Z} \\ \max\left(\left\lceil d\left(Z_2\right)\right\rceil, \lceil c_M \rceil, 0\right) & \text{otherwise.} \end{cases}$$

The conditional distribution used to find the confidence intervals is $f_Q\left(Z_1 | \boldsymbol{X}_{-1}\right) = f_Q\left(Z_1, \boldsymbol{X}_{-1}\right) / f_Q\left(\boldsymbol{X}_{-1}\right)$. Hence,

$$f_Q\left(Z_1 | \boldsymbol{X}_{-1}\right) = \mu^{-1} \left[s_M / \left(1 - s_M\right)\right]^{Z_1} \sum_{X_M \in D} \binom{n_{1M}}{X_M} \binom{n_2}{Z_1 - X_M}$$

where

$$\mu := \sum_{T=b}^{n_{1M}+n_2} \left[s_M / \left(1 - s_M\right)\right]^T \sum_{X_M \in D} \binom{n_{1M}}{X_M} \binom{n_2}{T - X_M}$$

is the normalising constant.

Acknowledgements

The authors would like to thank the two anonymous reviewers, whose comments greatly improved this research article.

Jack Bowden is funded by a Medical Research Council Methodology Research Fellowship; grant code MR/L012286/1. A. Toby Prevost was supported by the NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

The authors would also like to thank Fiona Walter for providing the data from the family history questionnaire study, which was funded by the National Institute for Health Research (NIHR) Research for Patient Benefit programme; grant reference number RfPB PB-PG-080713141.

References

1. Carmona FJ, Azuara D, Berenguer-Llergo A, Fernández AF, Biondo S, de Oca J, Rodriguez-Moranta F, Salazar R, Villanueva A, Fraga MF, Guardiola J, Capellá G, Esteller M, Moreno V. DNA methylation biomarkers for noninvasive diagnosis of colorectal cancer. *Cancer Prevention Research* 2013; **6**(7):656–665.
2. Madu CO, Lu Y. Novel diagnostic biomarkers for prostate cancer. *Journal of Cancer* 2010; **1**:150–177.

3. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 2001; **93**(14):1054–1061.
4. Gail MH, Costantino JP. Validating and improving models for projecting the absolute risk of breast cancer. *Journal of the National Cancer Institute* 2001; **93**(5):334–335.
5. Pepe MS, Feng Z, Longton G, Koopmeiners J. Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. *Statistics in Medicine* 2009; **28**:762–779.
6. Koopmeiners J, Feng Z, Pepe M. Conditional estimation after a two-stage diagnostic biomarker study that allows early termination for futility. *Statistics in Medicine* 2012; **31**(5):420–435.
7. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**(9):959–971.
8. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**(5):689–703.
9. Sampson AR, Sill MW. Drop-the-losers design: normal case. *Biometrical Journal* 2005; **47**(3):257–268.
10. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**(4):515–527.
11. Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multistage drop-the-losers trials. *Biometrical Journal* 2014; **56**(2):332–349.
12. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine* 2013; **32**(10):1677–1690.
13. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters* 1989; **8**(3):273–278.
14. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **24**(24):3697–3714.
15. van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**(6871):530–536.
16. Sill MW, Sampson AR. Drop-the-losers design: binomial case. *Computational Statistics & Data Analysis* 2009; **53**(3): 586–595.
17. Kimani PK, Todd S, Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. *Statistics in Medicine* 2013; **32**(17):2893–2910.
18. Tappin L. Unbiased estimation of the parameter of a selected binomial population. *Communications in Statistics - Theory and Methods* 1992; **21:4**:1067–1083.
19. Walter FM, Prevost AT, Birt L, Grehan N, Restarick K, Morris HC, Sutton S, Rose P, Downing S, Emery JD. Development and evaluation of a brief self-completed family history screening tool for common chronic disease prevention in primary care. *British Journal of General Practice* 2013; **63(611)**:e393–400.
20. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomials. *Biometrika* 1934; **26**(4):404–413.
21. Leisch F, Weingessel A, Hornik K. bindata: Generation of artificial binary data, 2012. http://CRAN.R-project.org/package=bindata, R package version 0.9-19.
22. Jovic G, Whitehead J. An exact method for analysis following a two-stage phase II cancer clinical trial. *Statistics in Medicine* 2010; **29**(30):3118–3125.
23. Koopmeiners JS, Vogel RI. Early termination of a two-stage study to develop and validate a panel of biomarkers. *Statistics in Medicine* 2013; **32**(6):1027–1037.
24. Zuber V, Strimmer K. Gene ranking and biomarker discovery under correlation. *Bioinformatics* 2009; **25**(20):2700–2707.
25. Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* 2004; **23**:881–896.

